Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

Comparative Evaluation Of Sparse, Dense, And Hybrid Retrieval Models On Indonesian Wikipedia

Tino Saputra*[1], Eric Julianto[2], Ari Widjonarko[3], Budi Tjahjono[4]

[1,2]Magister Computer,Computer Science, Esa Unggul University, Jakarta, Indonesia.
[3]Master Of Data Science, Data Science and AI, Monash University, Melbourne, Australia.
[4]Computer Science, Esa Unggul University, Jakarta, Indonesia.

Email: [1]saputra.tino85@gmail.com

Received : Apr 4, 2026; Revised : Apr 23, 2026; Accepted : Apr 26, 2026; Published : Jun 15, 2026

Abstract

This study presents a comparative evaluation of Information Retrieval (IR) models on the Indonesian Wikipedia corpus, focusing on sparse, dense, and hybrid retrieval approaches. The evaluated methods include TF-IDF and BM25 as sparse models, SBERT (MiniLM) as a dense retrieval model, and hybrid retrieval implemented through score fusion. The dataset consists of 713,044 Wikipedia articles, with experiments conducted using 1,000 test queries. Performance is measured using Precision@10 (P@10) and Mean Reciprocal Rank (MRR). The results show that BM25 achieves the highest performance, with a P@10 of 0.973 and an MRR of 0.9174, significantly outperforming TF-IDF and SBERT. Hybrid retrieval provides a slight performance improvement, where the BM25 + SBERT combination reaches a P@10 of 0.979 and an MRR of 0.9253 at higher α values. These findings indicate that lexical matching remains dominant in encyclopedic corpora, while semantic representations provide complementary improvements. However, the performance gain of hybrid retrieval is relatively marginal compared to the additional computational cost introduced by dense embedding and score fusion processes, indicating a trade-off between effectiveness and efficiency. These results highlight that, for low-resource languages such as Indonesian, lexical-based retrieval remains highly reliable, while hybrid approaches provide incremental improvements. Therefore, this study provides practical guidelines for developing efficient, scalable, and reliable Information Retrieval systems for Indonesian Wikipedia and other low-resource language corpora.

Keywords : *Information Retrieval, BM25, TF-IDF, Hybrid Retrieval, Wikipedia Indonesia*

1. INTRODUCTION

Information Retrieval (IR) systems are widely used in search engines, document repositories, and knowledge-based platforms such as Wikipedia. One of the main challenges in IR lies in how a system can effectively understand user queries and retrieve relevant documents, not only based on exact keyword matching but also on the underlying meaning of the query. Traditional IR approaches are largely dominated by sparse retrieval methods, which determine document relevance based on term matching between queries and documents. Among these, TF-IDF and BM25 are the most used baseline techniques. BM25 is known for its robustness in text retrieval tasks, as it considers both term frequency and document length, thereby reducing bias toward longer documents However, despite its effectiveness, sparse retrieval has limitations when queries and documents share similar meanings but use different vocabularies. In such cases, relevant documents may not be retrieved because lexical matching signals are not explicitly present.

To address this limitation, modern IR research has introduced dense retrieval approaches that leverage embedding representations to capture semantic similarity between queries and documents. Unlike sparse retrieval, dense retrieval does not rely solely on word overlap but instead measures

similarity within a semantic vector space. One of the most widely adopted methods for generating sentence embeddings is Sentence-BERT, which enables efficient similarity comparison between queries and documents [1]. Nevertheless, prior studies have shown that dense retrieval does not consistently outperform sparse methods, particularly when applied without domain-specific fine-tuning or in heterogeneous corpora, where embedding representations may fail to capture fine-grained lexical signals [2]. Its performance is highly dependent on factors such as corpus characteristics, dataset size, domain specificity, and document complexity [3]. This indicates that no single method can be considered universally optimal, and evaluation should always be aligned with the specific context and data characteristics.

Recent developments in IR have increasingly focused on hybrid retrieval approaches that combine the strengths of both sparse and dense methods. Sparse retrieval is effective in capturing exact keyword matches and specific terms, while dense retrieval excels at identifying semantic relationships when vocabulary differs. These two approaches can be integrated through score fusion techniques, where the relevance scores from each model are combined to produce more stable document rankings [4]. Hybrid retrieval has been successfully applied in various domains, including biomedical information systems and industrial FAQ applications, demonstrating improved retrieval performance compared to single-method approaches [5], [6], [7], [8]. Therefore, hybrid retrieval represents a promising direction for handling complex text corpora such as Wikipedia.

In the context of the Indonesian language, IR research is still predominantly based on sparse retrieval methods such as TF-IDF and BM25 due to their simplicity, efficiency, and relatively low computational requirements [9], [10]. At the same time, recent advancements in Indonesian Natural Language Processing (NLP) have shown significant progress through the adoption of transfer learning and modern representation-based models [11]. In addition, emerging studies have begun to explore the use of Large Language Models for evaluating Indonesian language systems, particularly in tasks such as entity linking [12]. Despite these developments, comprehensive studies that systematically compare sparse, dense, and hybrid retrieval methods specifically on Indonesian Wikipedia corpora remain limited. Considering that Indonesian Wikipedia provides a large-scale, diverse, and structurally rich dataset, it serves as an ideal benchmark for IR experimentation.

Based on these considerations, this study aims to conduct a comparative evaluation of multiple retrieval models on Indonesian Wikipedia, including TF-IDF and BM25 as sparse retrieval methods, SBERT/MiniLM as a dense retrieval approach, and a hybrid retrieval model based on score fusion. The evaluation is carried out using widely adopted IR metrics, namely Precision@10 and Mean Reciprocal Rank (MRR), allowing performance to be assessed both in terms of top-ranked relevance and overall ranking quality. This study is expected to provide insights into which retrieval approach is most suitable for Indonesian Wikipedia search tasks, as well as to examine whether hybrid retrieval can consistently deliver performance improvements over single-method approaches. Furthermore, the findings are intended to offer practical guidance for developing efficient and effective text-based search systems in the Indonesian context.

To systematically investigate these aspects, this study seeks to answer several key questions. First, how do sparse retrieval models (TF-IDF and BM25) perform compared to dense retrieval models (SBERT) on Indonesian Wikipedia? Second, to what extent do dense retrieval models lag behind strong lexical baselines such as BM25 when applied without domain-specific fine-tuning? Third, can hybrid retrieval approaches effectively improve retrieval performance by combining lexical and semantic representations? Finally, how does the fusion parameter (α) influence retrieval effectiveness in terms of Precision@10 and Mean Reciprocal Rank (MRR)?

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

2. METHOD

This study aims to develop and evaluate an Information Retrieval (IR) system on the Indonesian Wikipedia corpus. A quantitative approach is employed using a comparative experimental design, where multiple retrieval models are evaluated under the same testing conditions. The research workflow is structured systematically, starting from the literature study, dataset preparation, text preprocessing, and model development, including sparse retrieval (TF-IDF and BM25), dense retrieval (SBERT/MiniLM), and hybrid retrieval using score fusion. Each model is evaluated using identical experimental settings to ensure a fair and consistent comparison. The overall research pipeline is illustrated in Figure 1.

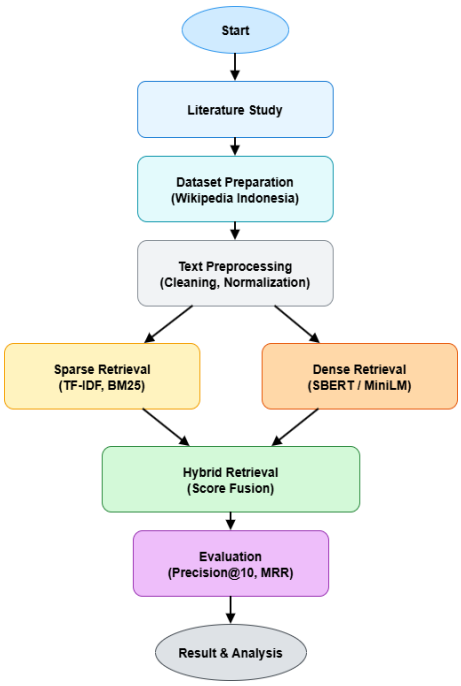


Figure 1. Overview of the Information Retrieval experiment pipeline.

The process starts from dataset preparation and preprocessing, followed by the implementation of sparse retrieval (TF-IDF, BM25) and dense retrieval (SBERT/MiniLM). The outputs of both approaches are combined using a hybrid retrieval method based on score fusion. Finally, the system performance is evaluated using Precision@10 (P@10) and Mean Reciprocal Rank (MRR).

2.1. Literature Study

Information Retrieval (IR) is a field that focuses on retrieving relevant documents based on user queries from large-scale text collections. The main objective of an IR system is not only to identify documents containing specific keywords, but also to rank the retrieved results according to their relevance to the user's information needs. As the volume of textual data continues to grow, IR methods have evolved from traditional keyword-matching approaches toward more advanced techniques that leverage semantic representations to better capture the meaning of both queries and documents[13], [14].

In general, IR approaches can be categorized into sparse retrieval and dense retrieval. Sparse retrieval determines relevance based on term matching between queries and documents, making its performance highly dependent on text representation and weighting schemes. Methods such as TF-IDF and BM25 are widely used as baselines due to their simplicity, efficiency, and relatively stable performance across different datasets, including document-based search systems in several studies conducted in Indonesia [9], [10]. In contrast, dense retrieval represents queries and documents using

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

vector embeddings in a semantic space, enabling the system to retrieve relevant documents even when there is no explicit term overlap. The development of dense retrieval has been significantly accelerated by Transformer-based models such as BERT and its variants, which have become the foundation of many modern retrieval architectures [13], [15].

In this study, related work is grouped into two categories. The first category consists of core literature that discusses the comparison of sparse, dense, and hybrid retrieval methods, including their evaluation benchmarks. A summary of this core literature is presented in Table 1 to illustrate the evolution of modern retrieval approaches and to support the methodological rationale for conducting a balanced comparison. The second category includes supporting literature that covers score fusion techniques, reproducibility aspects, and the development of NLP and IR in the Indonesian context. A summary of this supporting literature is presented in Table 2 to strengthen the methodological foundation and highlight the relevance of this study within the Indonesian research landscape.

Table 1. Summary of recent studies on sparse, dense, and hybrid retrieval.

Ref	Data/Domain	Approach	Key Method	Evaluation	Main Finding
[3]	Multi-domain	Sparse vs Dense	Strategy selection	P@k, MRR + efficiency	The trade-off between dense and sparse retrieval is influenced by corpus characteristics and computational cost..
[5]	Biomedical	Hybrid	Hybrid + reranking (zero-shot)	P@k, MRR	Hybrid approaches combined with reranking have been shown to be effective in literature search tasks.
[6]	Industrial FAQ	Hybrid	Dense-to-question + sparse-to-answer	Ranking metrics	Hybrid retrieval tends to provide more stable performance across different query variations and domains.
[13]	Survey	Dense IR	Neural IR survey	–	Dense retrieval has advanced rapidly, but still requires strong baselines for fair evaluation.
[16]	Benchmark	Sparse/Dense	BEIR (zero-shot)	MRR/nDCG	Heterogeneous benchmarks are important for testing the generalization capability of retrieval models
[17]	Commentary	Baseline fairness	Neural hype	–	Comparative evaluations should be balanced using established baselines such as BM25.
[18]	ODQA	Dense	Dense Passage Retrieval (DPR)	Top-k metrics	Dense retrievers are particularly effective for passage-level search.
[19]	General IR	Dense	ColBERT late interaction	Ranking metrics	Late interaction models offer an efficient and effective approach for ranking.
[20]	Evaluation track	Neural IR evaluation	TREC DL	MRR/nDCG	Neural ranking models are commonly evaluated using standardized competition benchmarks.

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

Table 2. Supporting studies on fusion, evaluation, Indonesian context, and retrieval foundations.

Ref	Focus	Contribution to This Study
[1]	SBERT	Provides the foundation for using sentence/document embeddings in similarity-based retrieval
[4]	Score fusion	Serves as the basis for combining sparse and dense scores in hybrid retrieval
[9]	IR literature (Indonesia)	Highlights the dominance of classical methods (TF-IDF/BM25) in local studies
[10]	Practical search engine (Indonesia)	Demonstrates real-world implementation of TF-IDF for document retrieval systems
[11]	IndoNLP transfer learning	Strengthens the context of NLP development in Indonesia for semantic models
[12]	Indonesian LLM evaluation	Indicates the growing use of LLMs for knowledge-based tasks in Indonesia
[15]	BERT foundation	Establishes the Transformer-based foundation for modern dense retrieval
[21]	Reproducibility	Emphasizes the importance of reproducible experiments in neural IR research
[22]	Lexical + semantic	Shows the effectiveness of combining lexical and semantic signals in retrieval
[23]	IR techniques (Indonesia)	Reinforces IR application and research trends in the Indonesian context
[24]	FAQ	Relevant to knowledge-based retrieval scenarios in the era of LLMs

2.2. Dataset

The dataset used in this study is derived from the Indonesian Wikipedia dump (idwiki), provided in XML format (idwiki-latest-pages-articles.xml.bz2). This corpus was selected due to its encyclopedic writing style, diverse topics, and relatively long document structures, making it suitable for Information Retrieval experiments. To accommodate computational constraints in the Google Colab environment, a representative subset of articles is utilized. In addition, a set of N = 1000 queries is constructed for evaluation purposes. This number is considered sufficient to produce stable and reliable evaluation results, while avoiding bias caused by a limited number of queries. This approach is consistent with common practices in modern IR evaluation benchmarks [16], [20].

The set of 1,000 queries used in this study was constructed by adapting Indonesian Wikipedia article titles and transforming them into natural language search queries. This approach aims to simulate realistic user search behavior while maintaining semantic alignment with the corresponding documents. Although the queries are not derived from real-world search logs, this controlled construction ensures consistency and reproducibility in the evaluation process.

2.3. Preprocessing

The preprocessing stage is conducted to normalize the text data before indexing and retrieval. This step follows standard procedures in Information Retrieval research, aiming to improve term matching quality for sparse retrieval methods such as TF-IDF and BM25, while maintaining consistent

input for dense retrieval models based on embeddings [9], [23]. After preprocessing, each document is stored in a clean text format and used as input for both sparse indexing and dense encoding processes.

2.4. Sparse Retrieval

Sparse retrieval is a traditional approach in Information Retrieval that determines document relevance based on term matching between queries and documents. In this approach, documents are represented as term-based vectors, and similarity is computed within the vector space.This method remains widely used as a baseline due to its computational efficiency, interpretability, and stable performance across various text collections, including long-form documents such as Wikipedia. In modern IR evaluations, strong sparse baselines are essential to ensure fair comparisons with neural-based approaches [17], [25].

In this study, two widely adopted methods are used: Term Frequency–Inverse Document Frequency (TF-IDF) and BM25. TF-IDF assigns higher weights to terms that frequently appear in a document but are rare across the corpus, as defined in Equation (1) [25].

$$TF\text{-}IDF(t,d) = TF(t,d) \times log\left(\frac{N}{DF(t)}\right) \quad (1)$$

where TF(t,d) represents the frequency of term t in document d, DF(t) denotes the number of documents containing term t, and N is the total number of documents in the collection. In practice, the relevance score between a query and a document is typically computed using cosine similarity in the TF-IDF vector space. This approach is widely used as a baseline due to its simplicity while still providing competitive performance in text retrieval tasks [9], [25].

In addition to TF-IDF, this study employs BM25 (2) [26], a probabilistic model designed to improve retrieval effectiveness by considering both term distribution and document length normalization. BM25 is regarded as one of the strongest baselines in sparse retrieval, as it effectively handles variations in document length commonly found in Wikipedia. In general, the BM25 score for a document d with respect to a query q is defined as follows:

$$BM25(d,q) = \sum_{t \in q} log\left(\frac{N - DF(t) + 0.5}{DF(t) + 0.5}\right) \cdot \frac{TF(t,d) \cdot (k_1 + 1)}{TF(t,d) + k_1\left(1 - b + b \cdot \frac{|d|}{avgdl}\right)} \quad (2)$$

where $|\ d\ |$ denotes the document length, avgdl represents the average document length, and k_1 and b are parameters that control the contribution of term frequency and document length normalization. BM25 is grounded in the probabilistic IR framework and remains a primary baseline in many modern retrieval studies due to its robustness and computational efficiency [3], [26], [27]. The BM25 model is implemented using standard parameter settings (k_1 = 1.5 and b = 0.75), following the default configuration commonly adopted in Information Retrieval libraries. These parameters are used without additional tuning to ensure consistency and a fair comparison across retrieval models.

Although recent IR research has increasingly focused on dense retrieval and neural ranking models, several studies emphasize that modern approaches should still be compared against strong sparse baselines to ensure the validity of the conclusions. This is supported by cross-domain evaluations showing that neural retrieval performance can vary and does not always consistently outperform traditional methods, particularly in heterogeneous document collections [28].

2.5. Dense Retrieval

Dense retrieval is a modern IR approach that represents both queries and documents as dense vector embeddings, allowing similarity to be measured based on semantic closeness rather than exact

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

term matching. This approach has advanced significantly with the development of Transformer-based models such as BERT [15].

In this study, dense retrieval is implemented using SBERT/MiniLM to generate embeddings for Indonesian Wikipedia documents. The dense retrieval model uses the *paraphrase-multilingual-MiniLM-L12-v2* architecture from Sentence-BERT, which produces 384-dimensional embeddings suitable for semantic similarity tasks. The similarity between queries and documents is computed using cosine similarity, as shown in Equation (3) [1]. Cosine similarity with query q and documents d is defined as follows:

$$sim(q,d) = \frac{\vec{q}\cdot\vec{d}}{\|\vec{q}\|\|\vec{d}\|} \quad (3)$$

Previous studies have shown that the performance of dense retrieval is influenced by domain generalization and representation learning strategies. Therefore, evaluation on a specific corpus remains necessary to validate model effectiveness [28]. Additionally, improvements in dense retrieval can be achieved through better sampling strategies and contrastive learning approaches The results obtained from dense retrieval are compared with sparse retrieval and further utilized in the hybrid retrieval approach.

2.6. Hybrid retrieval

Hybrid retrieval combines the strengths of sparse and dense retrieval approaches to produce more stable and accurate ranking results. Sparse retrieval excels in capturing exact keyword matches, while dense retrieval is effective in understanding semantic relationships when vocabulary differences occur [3], [6]. In this study, hybrid retrieval is implemented using a score fusion approach by combining BM25 scores and SBERT/MiniLM scores. The score combination is expressed in Equation (4):

$$Score_{hybrid}(d,q) = \alpha \cdot BM25(d,q) + (1-\alpha)\cdot sim(q,d) \quad (4)$$

with $\alpha \in [0,1]$ as a weighting parameter that controls the contribution of sparse and dense scores. The fusion parameter α is evaluated using multiple values (0.2, 0.4, 0.6, and 0.8) to analyze its effect on retrieval performance and to determine the optimal balance between lexical (BM25) and semantic (SBERT) contributions.

Such fusion strategies are commonly used in IR systems to improve ranking performance by leveraging complementary strengths of different models [4]. Studies in domains such as biomedical and industrial FAQ systems have also shown that hybrid retrieval can improve effectiveness compared to single-method approaches, particularly for diverse query variations [5], [6]. In this study, the value of α is tested across several configurations to identify the combination that yields the best performance based on the evaluation metrics.

2.7. Evaluation

The evaluation is conducted to measure the effectiveness of the retrieval models in returning relevant documents at the top ranks. In this study, two primary metrics are used: Precision@10 (P@10) and Mean Reciprocal Rank (MRR), as both are widely adopted in modern retrieval evaluation and effectively represent the quality of top-ranked search results from the user's perspective [16], [20]. Precision@10 measures the proportion of relevant documents within the top 10 retrieved results. It is defined in Equation (5) as follows:

$$P@10 = \frac{|Rel@10|}{10} \quad (5)$$

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

With $|Rel@10|$ represents the number of relevant documents appearing within the top 10 ranked results. Meanwhile, MRR measures the ranking quality based on the position of the first relevant document, meaning that the earlier a relevant document appears, the higher the score. MRR is defined in Equation (6) as follows:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (6)$$

With $|Q|$ where N denotes the number of evaluation queries, and rank_i represents the position of the first relevant document for the i-th query. All methods (TF-IDF, BM25, SBERT, and hybrid retrieval) are evaluated using the same set of queries to ensure a fair and consistent comparison across models.

3. RESULT

3.1. Eksperimnt and Setup

The dataset used in this study is derived from the Indonesian Wikipedia dump, which is extracted into a collection of text documents. Based on the parsing results, a total of 713,044 articles were successfully obtained, where each document consists of a title and the corresponding article content. Since the raw Wikipedia data still contains internal markup (such as templates, categories, media files, and link formatting), the extracted documents cannot be directly used for retrieval tasks without prior cleaning, as this noise may affect the quality of text representation.

The preprocessing stage is carried out through two main steps: markup cleaning and text normalization. Markup cleaning is performed to remove templates ({{...}}), media elements ([[File:...]]), categories ([[Category:...]]), as well as HTML tags, while converting internal Wikipedia links into plain text. Subsequently, normalization is applied by converting all characters to lowercase and removing non-alphanumeric symbols to produce a more consistent textual representation. To ensure stable processing under the memory constraints of the Google Colab environment, the preprocessing is executed using a batch processing strategy with a chunk size of 20,000 documents per iteration. The final output of this stage is a cleaned dataset containing the fields "title" and "clean", which are then used as inputs for indexing and retrieval experiments.

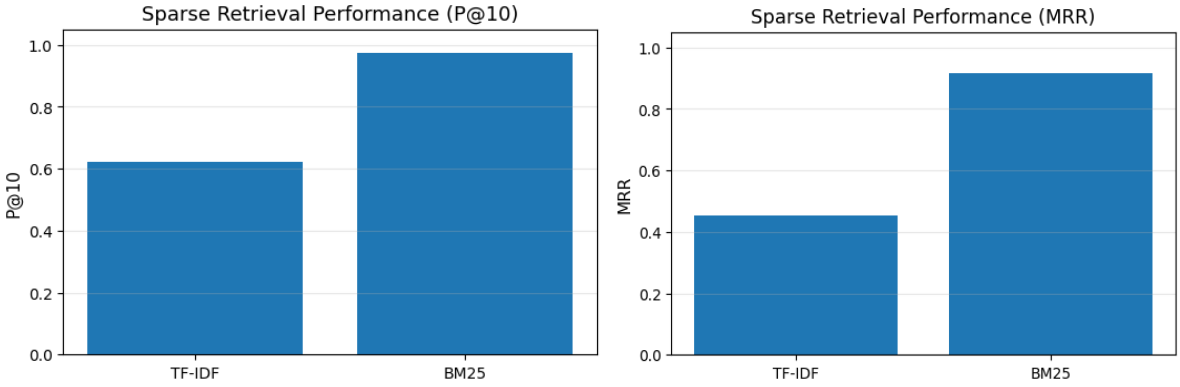
3.2. Sparse Retrieval (TF-IDF dan BM25)

In the sparse retrieval stage, two classical models are evaluated: TF-IDF and BM25. The evaluation is conducted using 1,000 test queries, with Precision@10 (P@10) and Mean Reciprocal Rank (MRR) as the evaluation metrics. The results show that BM25 significantly outperforms TF-IDF. TF-IDF achieves a P@10 of 0.622 and an MRR of 0.4549, while BM25 reaches a P@10 of 0.973 and an MRR of 0.9174. This substantial difference indicates that BM25 is more effective in ranking relevant documents at the top positions in the context of Indonesian Wikipedia search. This performance advantage can be attributed to BM25's weighting mechanism, which considers both term frequency and document length normalization. Such characteristics make BM25 more suitable for handling encyclopedic documents with varying lengths, as commonly found in Wikipedia. These findings reinforce the role of BM25 as a strong baseline in modern Information Retrieval, particularly for large-scale and heterogeneous text collections.

From an interpretative perspective, TF-IDF tends to be sensitive to the occurrence of common terms and lacks a balancing mechanism as robust as BM25 in handling variations in document length. In contrast, BM25 is more stable in ranking truly relevant documents at the top positions. These findings reinforce that BM25-based sparse retrieval remains a highly competitive baseline for text-based search

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

systems, particularly when dealing with documents that are long and rich in vocabulary, such as Wikipedia.

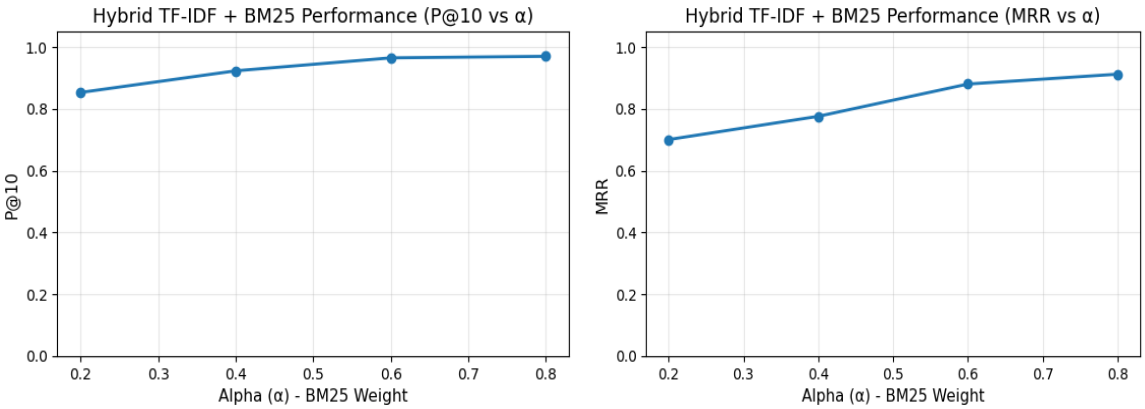


(a) (b)

Figure 2. Evaluation results of sparse retrieval: (a) P@10 evaluation and (b) MRR evaluation.

3.3. Hybrid Retrieval (TF-IDF + BM25 Score Fusion)

he experiment is further extended using a hybrid approach based on score fusion between TF-IDF and BM25. The scores from both methods are first normalized using min–max scaling to ensure they are within a comparable range, and then combined using a weighting parameter α to control the contribution of BM25 to the final score. In this experiment, several α values are evaluated, namely 0.2, 0.4, 0.6, and 0.8. The evaluation results show a consistent improvement in performance as the contribution of BM25 in the score fusion increases.



(a) (b)

Figure 3. Evaluation results of hybrid retrieval: (a) P@10 evaluation and (b) MRR evaluation

Based on the graph, the hybrid model with α = 0.2 achieves a P@10 of 0.8530 and an MRR of 0.7003, while α = 0.4 yields a P@10 of 0.9230 and an MRR of 0.7757. A more substantial improvement is observed at α = 0.6, with a P@10 of 0.9650 and an MRR of 0.8802. The best performance is obtained at α = 0.8, reaching a P@10 of 0.9700 and an MRR of 0.9121, indicating that the hybrid approach performs close to pure BM25. These findings suggest that, in the context of the Indonesian Wikipedia corpus, lexical signals from BM25 remain the dominant factor, while TF-IDF acts as a complementary component, particularly in cases where term distribution is better captured by simpler vector representations. Hasil Dense Retrieval (SBERT /MiniLM

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

After evaluating the sparse and hybrid (sparse-based) approaches, the experiment is extended to dense retrieval using semantic representations with the Sentence-BERT (SBERT) MiniLM model. In this approach, each document in the Indonesian Wikipedia corpus is converted into a 384-dimensional embedding vector using the SBERT encoder. Retrieval is then performed by measuring the similarity between query embeddings and document embeddings using cosine similarity. The evaluation is conducted using the same experimental setup as in the previous stages, namely 1,000 test queries with Precision@10 (P@10) and Mean Reciprocal Rank (MRR) as the evaluation metrics, to ensure a consistent comparison across models.

The results indicate that the performance of SBERT-based dense retrieval is still lower than that of the sparse models. SBERT achieves a P@10 of 0.325 and an MRR of 0.2431. These results suggest that the embedding-based model is not yet able to consistently rank relevant documents at the top positions in the context of Indonesian Wikipedia search. In general, this limitation can be attributed to the characteristics of Wikipedia, which consists of long and highly diverse documents. As a result, a single document-level embedding may fail to capture important fine-grained information. Furthermore, the SBERT model used in this study is a general-purpose model without domain-specific fine-tuning on the Indonesian Wikipedia corpus, which may lead to suboptimal semantic alignment between queries and documents compared to direct lexical matching methods such as BM25.

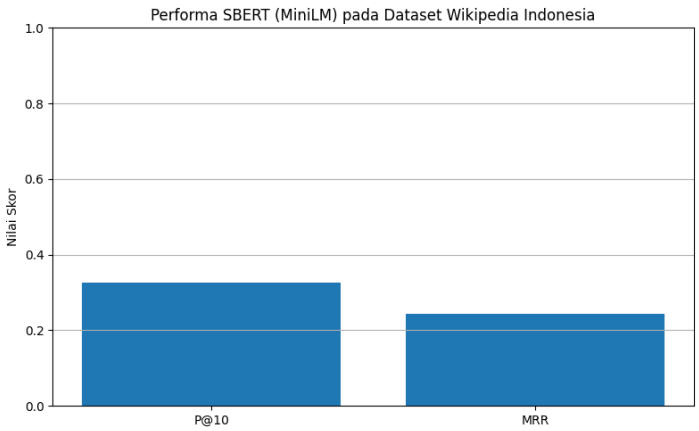


Figure 4. Evaluation Dense retrieval P@10, MRR

This experiment demonstrates that dense retrieval does not always outperform sparse retrieval, particularly when dealing with encyclopedic corpora and queries that are explicitly term-based. Therefore, these findings provide a strong foundation for extending the study to a hybrid approach that combines the lexical signals of BM25 with the semantic representations of SBERT, with the expectation that dense retrieval can improve specific cases that are not effectively captured by purely sparse methods.

3.4. Hybrid Retrieval (BM25 + SBERT Score Fusion)

The experiment is then extended using a hybrid retrieval approach that combines the strengths of lexical signals from BM25 and semantic representations from SBERT. In this scheme, both BM25 scores and SBERT similarity scores are first normalized using min–max scaling to ensure they are within a comparable range. The scores are then fused using a weighting parameter α, which controls the contribution of BM25 to the final score. The fusion formula follows a linear combination: $\alpha \cdot \text{BM25} + (1-\alpha) \cdot \text{SBERT}$, where a larger α indicates greater reliance on BM25, while a smaller α assigns more weight to SBERT. In this experiment, three α values are evaluated, namely 0.2, 0.5, and 0.8, to analyze the impact of each component on retrieval performance.

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

The evaluation results show that the BM25+SBERT hybrid approach significantly improves performance compared to pure dense retrieval and also outperforms the TF-IDF baseline. With α = 0.2, the hybrid model achieves a P@10 of 0.628 and an MRR of 0.5383. This result already surpasses TF-IDF (P@10 = 0.622; MRR = 0.4549), although it remains considerably lower than pure BM25. When the BM25 contribution is increased to α = 0.5, performance improves substantially, reaching a P@10 of 0.979 and an MRR of 0.9214. The best performance is obtained at α = 0.8, with a P@10 of 0.979 and an MRR of 0.9253. These findings indicate that BM25 remains the dominant factor in Indonesian Wikipedia retrieval, while the addition of semantic signals from SBERT provides a small but consistent improvement in ranking quality, particularly in terms of MRR.

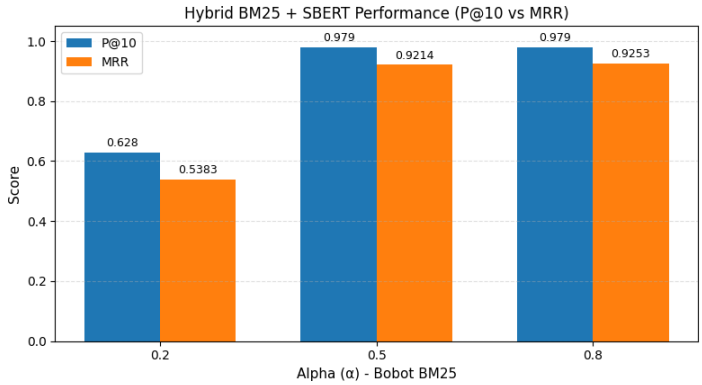


Figure 5. Hybrid retrieval BM25+SBERT dengan Evaluasi P@10 & MRR

From an interpretative perspective, the BM25+SBERT hybrid approach performs better because it preserves the lexical precision of BM25 in capturing explicit query terms, while leveraging SBERT embeddings as a semantic enhancer to handle variations in wording or similarity in meaning that are not directly captured by term matching. However, since the Indonesian Wikipedia corpus is rich in terminology and many queries exhibit strong lexical overlap, the contribution of SBERT does not always lead to substantial improvements. This is reflected in the observation that the best hybrid performance only slightly surpasses BM25 in terms of MRR. Therefore, SBERT primarily serves as a complementary component that enhances ranking quality in specific cases, rather than replacing the dominance of BM25. In this context, the hybrid approach can be considered a more stable compromise, particularly for retrieval systems that aim to balance lexical accuracy with semantic understanding.

4. DISCUSSIONS

Based on the overall evaluation on the Indonesian Wikipedia corpus, the sparse retrieval approach remains the most stable and dominant. The BM25 model achieves the highest P@10 compared to TF-IDF and SBERT, indicating that lexical matching is still highly effective for encyclopedic documents with long structures and rich vocabularies. This result suggests that Indonesian Wikipedia articles tend to contain explicit keyword overlap between queries and relevant documents, making term-based retrieval highly reliable. In contrast, SBERT (MiniLM), as a pure dense retrieval approach, yields significantly lower P@10 scores, indicating that semantic representations alone are not yet sufficiently robust to replace lexical signals in this dataset context.

This finding is consistent with prior benchmark studies such as BEIR, where dense retrieval models often fail to outperform strong lexical baselines in zero-shot settings. Without domain-specific fine-tuning, dense models rely on general semantic representations that may not align well with the structure and characteristics of encyclopedic corpora. In the case of Indonesian Wikipedia, the mismatch is further amplified by linguistic variation and limited training exposure in low-resource settings. As a

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

result, SBERT embeddings may capture general topic similarity but fail to identify fine-grained relevance required for top-ranked retrieval.

Another important factor contributing to the poor performance of SBERT is the granularity mismatch between document representation and retrieval requirements. In this study, dense retrieval is applied at the document level, where each Wikipedia article is represented as a single embedding vector. However, Wikipedia documents are typically long and contain multiple subtopics, which cannot be fully represented within a single embedding. In addition, SBERT models generally operate with a maximum input length (typically around 512 tokens), leading to truncation or loss of important contextual information. This limitation significantly reduces the model's ability to rank relevant documents accurately at top positions.

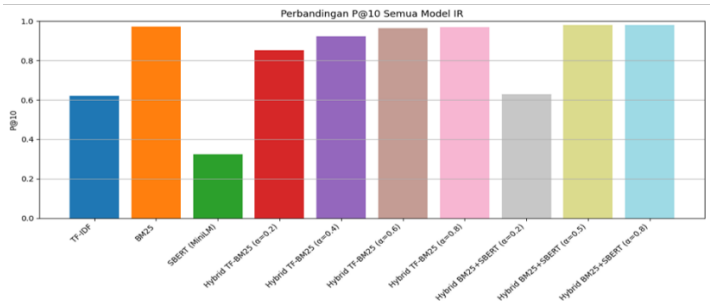


Figure 6. P@10 Comparison of All Information Retrieval Models

In the hybrid retrieval setting, score fusion provides a clear improvement over certain single-method approaches, particularly when the contribution of BM25 is dominant. The TF-IDF + BM25 hybrid shows an increasing performance trend as α increases, while the BM25 + SBERT hybrid achieves optimal performance at medium to high α values. This indicates that semantic representations from SBERT provide complementary signals, but their contribution remains secondary to lexical matching. In other words, hybrid retrieval improves robustness rather than fundamentally replacing sparse retrieval.

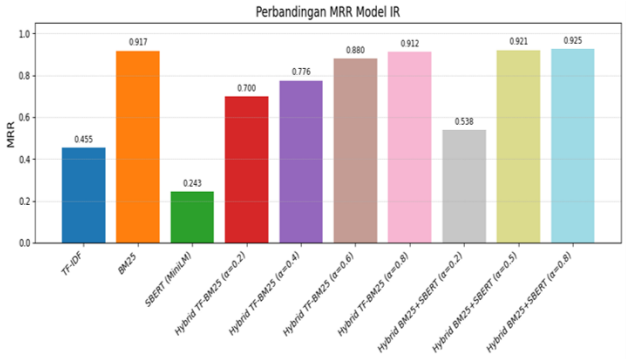


Figure 7. Comparison MRR All Model IR

In terms of MRR, the performance across models shows a trend consistent with P@10, but with a stronger emphasis on ranking quality at top positions. BM25 remains the most stable model, achieving the highest MRR among single-method approaches, while SBERT yields the lowest score. This further confirms that dense retrieval is not yet effective in prioritizing highly relevant documents in Indonesian Wikipedia search. Within the hybrid setting, a significant improvement in MRR is observed as the contribution of BM25 increases, particularly in the BM25 + SBERT combination. The best performance

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

is achieved at higher α values, reinforcing the conclusion that lexical signals remain dominant in determining ranking quality. From a practical perspective, these findings have important implications for Indonesian NLP applications. While dense retrieval and neural search architectures are increasingly popular in global research, their direct adoption in Indonesian document search systems may still be premature without sufficient domain-specific fine-tuning. In contrast, BM25 offers a more reliable, computationally efficient, and interpretable solution for large-scale document retrieval in Indonesian. Therefore, for many real-world applications such as knowledge management systems, digital archives, and enterprise search, lexical-based retrieval remains a strong baseline and often the most practical choice.

Despite these contributions, several limitations should be acknowledged. First, the dense retrieval model used in this study is a general-purpose multilingual SBERT model that is not specifically fine-tuned for Indonesian retrieval tasks. Second, the model is constrained by input length limitations (approximately 512 tokens), which may lead to information loss when encoding long Wikipedia articles. Third, this study applies document-level retrieval rather than passage-level or chunk-based retrieval, which may limit the ability of dense models to capture fine-grained relevance. Finally, the hybrid approach is limited to score fusion and does not explore more advanced architectures such as late interaction or reranking models. Future research may address these limitations by incorporating passage-level retrieval, domain-specific fine-tuning, and more advanced hybrid architectures. These approaches may help unlock the full potential of dense and hybrid retrieval models for Indonesian language applications.

5. CONCLUSION

This study presents a comprehensive evaluation of sparse, dense, and hybrid retrieval models on the Indonesian Wikipedia corpus. The results demonstrate that BM25 consistently outperforms TF-IDF and SBERT (MiniLM) in both Precision@10 and Mean Reciprocal Rank, confirming that lexical matching remains highly effective for long and information-rich encyclopedic documents. While hybrid retrieval approaches provide measurable improvements over certain configurations, the gains remain relatively modest and are strongly influenced by the dominance of the BM25 component. These findings indicate that, in the context of Indonesian Wikipedia, semantic representations alone are not yet sufficient to replace strong lexical baselines, particularly in zero-shot settings without domain-specific fine-tuning.

From a broader perspective, this study highlights important implications for Informatics and Natural Language Processing applications in developing language contexts such as Indonesian. The results suggest that the direct adoption of dense retrieval or neural search architectures may still be premature for large-scale document retrieval tasks without substantial adaptation and training resources. Instead, BM25 remains a reliable, efficient, and practical solution for real-world systems involving knowledge repositories and document search. Future research may explore more advanced retrieval architectures, such as cross-encoder reranking models to improve ranking precision, as well as passage-level retrieval through document chunking to better handle long texts. Additionally, domain-specific fine-tuning of dense models and the integration of hybrid retrieval with reranking strategies may further enhance performance for Indonesian Information Retrieval systems.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

AUTHOR INFORMATION

Tino Saputra acted as the primary and corresponding author, responsible for study design, implementation, experiments, data analysis, manuscript writing, and submission. Eric contributed by collecting and organizing relevant references to support the research. Ari provided external academic support and feedback from an international campus perspective. Budi Tjahjono supervised the study and provided scientific guidance, review, and final revisions to improve the manuscript.

REFERENCES

[1] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019. doi: 10.18653/v1/D19-1410.

[2] G. Sidiropoulos, N. Voskarides, S. Vakulenko, and E. Kanoulas, "Combining Lexical and Dense Retrieval for Computationally Efficient Multi-hop Question Answering Georgios," *Assoc. Comput. Linguist.*, pp. 1–6, 2021, doi: 10.18653/v1/2021.sustainlp-1.7.

[3] N. Arabzadeh, H. Zamani, and others, "Predicting Efficiency/Effectiveness Trade-offs for Dense vs. Sparse Retrieval Strategy Selection," in *Proceedings of the 2021 ACM International Conference on Information and Knowledge Management (CIKM)*, 2021. doi: 10.1145/3459637.3482159.

[4] D. Metzler, W. B. Croft, and A. M. Diaz, "Combining Different Retrieval Models Using Score Fusion," *Inf. Retr. Boston.*, vol. 12, no. 1, pp. 1–27, 2009, doi: 10.1007/s10791-008-9061-6.

[5] J. Lu, J. Ma, and K. Hall, "Zero-shot Hybrid Retrieval and Reranking Models for Biomedical Literature," in *CEUR Workshop Proceedings*, 2022.

[6] J. Seo, "Dense-to-Question and Sparse-to-Answer: Hybrid Retriever System for Industrial Frequently Asked Questions," *Mathematics*, vol. 10, no. 8, p. 1335, 2022.

[7] Z. Tu and S. J. Padmanabhan, "MIA 2022 Shared Task Submission : Leveraging Entity Representations , Dense-Sparse Hybrids , and Fusion-in-Decoder for Cross-Lingual Question Answering," *Assoc. Comput. Linguist.*, vol. Proceeding, pp. 100–107, 2022, doi: 10.18653/v1/2022.mia-1.10.

[8] L. Xu, Z. Su, M. Yu, J. Li, F. Meng, and Z. Jie, "Dense Retrievers Can Fail on Simple Queries : Revealing The Granularity Dilemma of Embeddings," *Assoc. Comput. Linguist.*, pp. 19295–19305, 2025, doi: 10.18653/v1/2025.findings-emnlp.1051.

[9] E. N. Azizah and A. N. Handayani, "Permodelan pada Information Retrieval: Literature Review," *J. Inov. Teknol. dan Edukasi Tek.*, vol. 2, no. 11, pp. 527–535, 2022, doi: 10.17977/um068v2i112022p527-535.

[10] R. Rudiansyah, R. Wahyuni, and M. Andri, "Search Engine Menggunakan Metode Information Retrieval," *J. SANTI*, vol. 2, no. 1, pp. 21–30, 2022.

[11] A. S. Ekakristi, A. F. Wicaksono, and R. Mahendra, "Intermediate-task Transfer Learning for Indonesian NLP Tasks," *Nat. Lang. Process. J.*, 2025, doi: 10.1016/j.nlp.2025.100161.

[12] R. H. Gusmita, A. F. Firmansyah, H. M. Zahera, and A.-C. Ngonga Ngomo, "ELEVATE-ID: Extending Large Language Models for End-to-End Entity Linking Evaluation in Indonesian," *Data Knowl. Eng.*, vol. 161, 2026, doi: 10.1016/j.datak.2025.102504.

[13] H. Zamani, M. Dehghani, W. B. Croft, and M. Bendersky, "Neural Information Retrieval: A Survey," *ACM Trans. Inf. Syst.*, vol. 40, no. 2, 2022, doi: 10.1145/3486250.

[14] Z. Xu, Z. Dou, J.-R. Wen, and R. Zhang, "A Survey of Model Architectures in Information Retrieval," *arXiv Prepr.*, 2025.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019. doi: 10.18653/v1/N19-1423.

[16] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," 2021. doi: 10.48550/arXiv.2104.08663.

[17] J. Lin, "The Neural Hype and Comparisons Against Weak Baselines," *SIGIR Forum*, vol. 52,

Vol.7, No. 3, June 206, Page. 2920-2934
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5776

no. 2, pp. 40–51, 2019, doi: 10.1145/3308774.3308778.

[18] V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

[19] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in *Proceedings of the International ACM SIGIR Conference*, 2020. doi: 10.1145/3397271.3401075.

[20] N. Craswell, B. Mitra, E. Yilmaz, and D. Campos, "Overview of the TREC 2020 Deep Learning Track," in *Proceedings of TREC*, 2020.

[21] S. Marchesin, A. Purpura, F. Silvestri, R. Perego, and G. Faggioli, "Focal Elements of Neural Information Retrieval Models: An Outlook through a Reproducibility Study," *Inf. Process. \& Manag.*, 2020, doi: 10.1016/j.ipm.2020.102201.

[22] L. Gao, Z. Dai, T. Chen, Z. Fan, B. Van Durme, and J. Callan, "Complement Lexical Retrieval Model with Semantic Residual Embeddings," in *Advances in Information Retrieval (ECIR 2021)*, 2021. doi: 10.1007/978-3-030-72240-1_11.

[23] I. G. N. A. Jayarana, I. G. W. Darma, I. W. A. Juliantara, and I. M. A. W. Putra, "Study Literatur Information Retrieval Model: Teknik dan Aplikasi," *J. Sutasoma*, vol. 3, no. 1, pp. 61–69, 2025, doi: 10.58878/sutasoma.v3i2.392.

[24] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[25] C. D. Manning, P. Raghavan, and H. Schuetze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[26] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[27] K. Sparck Jones, S. Walker, and S. E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments," *Inf. Process. Manag.*, vol. 36, no. 6, pp. 779–808, 2000, doi: 10.1016/S0306-4573(00)00015-7.

[28] S. MacAvaney, A. Yates, A. Cohan, and N. Goharian, "On the Robustness of Neural Ranking Models Across Domains," *ACM Trans. Inf. Syst.*, 2022, doi: 10.1145/3512345.